Effects of a universal parenting program for highly adherent parents: A propensity score matching approach

Eisner, M P ; Nagin, D ; Ribeaud, D ; Malti, T

Abstract: This paper examines the effectiveness of a group-based universal parent training program as a strategy to improve parenting practices and prevent child problem behavior. In a dissemination trial, 56 schools were first selected through a stratified sampling procedure, and then randomly allocated to treatment conditions. 819 parents of year 1 primary school children in 28 schools were offered Triple P. 856 families in 28 schools were allocated to the control condition. Teacher, primary caregiver and child self-report data were collected at baseline, post, and two follow-up assessments. Analyses were constrained to highly adherent parents who completed all four units of the parenting program. A propensity score matching approach was used to compare parents fully exposed to the intervention with parents in the control condition, who were matched on 54 baseline characteristics. Results suggest that the intervention had no consistent effects on either five dimensions of parenting practices or five dimensions of child problem behavior, assessed by three different informants. These findings diverge from findings reported by program developers and distributors. Potential explanations for the discrepancy and implications for future research are discussed.

DOI: https://doi.org/10.1007/s11121-011-0266-x

Prev Sci
DOI 10.1007/s11121-011-0266-x

Effects of a Universal Parenting Program for Highly Adherent Parents: A Propensity Score Matching Approach

Manuel Eisner · Daniel Nagin · Denis Ribeaud · Tina Malti

Abstract This paper examines the effectiveness of a group-based universal parent training program as a strategy to improve parenting practices and prevent child problem behavior. In a dissemination trial, 56 schools were first selected through a stratified sampling procedure, and then randomly allocated to treatment conditions. 819 parents of year 1 primary school children in 28 schools were offered Triple P. 856 families in 28 schools were allocated to the control condition. Teacher, primary caregiver and child self-report data were collected at baseline, post, and two follow-up assessments. Analyses were constrained to highly adherent parents who completed all four units of the parenting program. A propensity score matching approach was used to compare parents fully exposed to the intervention with parents in the control condition, who were matched on 54 baseline characteristics. Results suggest that the intervention had no consistent effects on either five dimensions of parenting practices or five dimensions of child problem behavior, assessed by three different informants. These findings diverge from findings reported by program developers and distributors. Potential explanations for the discrepancy and implications for future research are discussed.

Keywords Prevention · Randomized controlled trial · Propensity score matching · Parent training

M. Eisner (✉)
Institute of Criminology, University of Cambridge,
Cambridge, UK
e-mail: mpe23@cam.ac.uk

D. Nagin
Heinz College, Carnegie Mellon University,
Pittsburgh, PA, USA

D. Ribeaud
Department of Sociology, Swiss Federal Institute of Technology,
Zurich, Switzerland

T. Malti
Department of Psychology, University of Toronto,
Mississauga, ON, Canada

Published online: 10 January 2012

Problematic parenting such as harsh and inconsistent discipline, low involvement, and poor supervision are major predictors of antisocial behavior in children and adolescents (e.g., Capaldi et al. 1997; Loeber and Stouthamer-Loeber 1986). Accordingly, interventions that aim at changing parenting behavior are seen as a key strategy for reducing and preventing child problem behavior. Various meta-analyses conclude that parent training has a positive influence on parenting practices and child problem behavior (e.g., Lundahl et al. 2006; Maughan et al. 2005; Piquero et al. 2009; Reyno and McGrath 2006; Serketich and Dumas 1996). Therefore, training programs have gained popularity not only for treating disfunctional families, but also for the community-based early prevention of child and adolescent problem behaviors (Sanders et al. 2003; Spoth et al. 2002).

However, the available evidence base still raises questions. Thus, evidence for positive effects is strongest for indicated treatment in clinical settings, while findings are less unequivocal for parent training as a community-based preventative approach. For example, Spoth (2001), McTaggart and Sanders (2003), and Gross et al. (2009) report positive effects on child problem behaviors, but Gottfredson et al. (2006) and Hiscock et al. (2008) found no effects. Second, many results come from studies with small samples, a tight control over treatment delivery and measures of short-term effects only. Yet several meta-analyses find that effect sizes decrease in studies with large Ns (e.g., Farrington and Welsh 2007; Piquero et al. 2009)

and in studies that report effects for follow-up measures (Lundahl et al. 2006). These findings raise doubts about whether average effects can be generalized to population-wide prevention that aims at having long-term impact. Finally, evidence suggests that independent evaluations report, on average, lower effect sizes than studies conducted by the developers or distributors of a treatment (Friedman and Richter 2004; Perlis et al. 2005; Petrosino and Soydan 2005). In prevention research, failed attempts to replicate findings from developer-led studies include substance abuse programs (e.g., Hallfors et al. 2006; St. Pierre et al. 2005), anti-bullying programs (e.g., Bauer et al. 2007; Jenson and Dieterich 2007) and parenting programs (e.g., Gottfredson et al. 2006). Yet successful independent replication is essential for establishing effectiveness outside the controlled environment of developer-led trials.

In this paper we report findings from an independent dissemination trial of a group-based parent training program offered as a universal preventive intervention. The tested program is Triple P, a program found to be effective in numerous studies conducted by the program developers and by distributors in several countries (Nowak and Heinrichs 2008). In a previous study, Malti et al. (2011) reported findings based on an intention-to-treat basis. In this paper we limit the analyses to *highly adherent parents* who fully completed the program. To derive unbiased estimates of treatment effects we use propensity score matching, a statistical approach developed to estimate treatment effects on the treated when self-selection into treatment occurs.

The Study

The data derive from the Zurich Project on the Social Development of Children (*z-proso),* a prospective, longitudinal study of children that entered 1 of 56 primary schools in the City of Zurich, Switzerland, in the year 2004 (Eisner and Ribeaud 2005). Embedded in the longitudinal study, the School Department of Zurich implemented two prevention programs, namely the family-based parenting skills program *Triple P* (Positive Parenting Program; e.g., Sanders 1999), and the school-based social skills program *PATHS* (Promoting Alternative Thinking Strategies; e.g., Greenberg et al. 1998).

The units for sampling and treatment allocation were primary schools. Schools rather than individuals or classes were randomly allocated to treatment conditions in order to minimize possible contamination, or spillover effects between treatment conditions. The sampling frame for the study was formed by all 90 public primary schools in the City of Zurich. Schools were first stratified by school size and socio-economic background of the school district. Then a stratified sample of 56 schools was drawn, comprising 1675 first year primary school children. All selected schools participated in the study. Due to the stratified sampling procedure, the 56 schools formed 14 "quadruplets" of schools. Each quadruplet comprised four schools of similar size and socio economic background of the catchment area.

Subsequent to the formation of the sample, a 2 × 2 factorial design was used to randomly allocate schools in each quadruplet to four treatment conditions: PATHS only, Triple P only, PATHS and Triple P combined, and control group. The parent training program was implemented between waves 1 and 2 of the longitudinal study, while the core of the school-based social skills program was implemented between waves 2 and 3 (i.e., during year 2 of primary school).

The Intervention

Triple P was developed in Australia by Sanders and colleagues as a parenting and family support strategy that comprises varying levels of intensity (Sanders 1999; Sanders et al. 2002, 2003). It is amongst the most thoroughly evaluated parent training programs in the world. A meta-analysis by Nowak and Heinrichs (2008) identified 55 studies that had assessed the effectiveness of Triple P on a variety of outcome measures. The study reports significant positive effects on parenting (Cohen's d=0.38), child problem behavior (d=0.35), and parental well-being (d=0.17).

In the current study, level 4 Triple P, also known as *Standard Triple P*, was implemented. Its core element is a course that comprises four units of 2 to 2.5 h, delivered in a group format. The units address themes such as positive parenting, techniques to support desired behaviors, routines that help to avoid the escalation of conflicts, or planning ahead. To support active learning, units comprise video clips, group discussion, role play and homework for the parents. Additionally, the program includes up to four follow-up telephone contacts, conducted by the course providers, of 15–30 min with each participant.

The implementation team of the school authority managed the recruitment and organization of the courses. The target group comprised all parents of first grade children in the 28 schools allocated to the Triple P condition. In October 2004 the schools sent an information package to the parents. Also, experienced Triple P providers introduced Triple P during the first parent-teacher meetings of grade 1.

Participation was free of costs. Courses were offered in all school districts and travel distances were generally below one mile. The program was offered at different weekdays and times of the day, and a free child-care service was available to all participants. Additional efforts were made to recruit families with an immigrant background: The information package was translated into the nine most important languages of immigrant minorities. Also, *Triple P International* translated the program into Albanian, Portuguese and

Turkish. In Zurich, these three languages are spoken by sizeable immigrant minorities who experience, on average, a considerable extent of social disadvantage (Eisner and Ribeaud 2007). Courses were delivered by licensed providers selected in collaboration with Triple P Switzerland. German-speaking providers had previous experience in delivering the courses. For the Albanian, Turkish and Portuguese programs new providers were recruited by the implementation team and trained by *Triple P Switzerland.*

The implementation team organized 41 Triple P courses. Thirty-three courses were held in German, 3 in Turkish, 2 each in Portuguese and Albanian and 1 in English. Courses began in May 2005, about 6 months after the median date of the baseline parent interviews. They were completed in early July 2005. Parents of 257 children enrolled for the program (31.3% of the target population). Parents of 220 children (26.8%) attended at least one session. Parents of 153 children (18.6%) completed all four course units. One hundred forty-four of these the parents participated in wave 1 of the longitudinal study, meaning that background information is available. Eisner et al. (2011) examined determinants of parental engagement. Results suggest that parents who engaged with the program were more likely to come from breadwinner families, to be Swiss, to have a high socio economic background, to have previously used parent services and to be highly integrated in neighborhood social networks. However, program compliers did not differ from non-compliers in respect of levels of parenting problems or child problem behaviors.

The program was delivered to high standards. Participant overall satisfaction with the program was 4.33 (SD=0.89) and provider competency was rated at 4.65 (SD=0.73) on a five-point scale. Course providers estimated that 93% of the full course material was delivered during the sessions.

The Longitudinal Study

We use data from the first four waves of the longitudinal study. Waves 1 to 3 were conducted at annual intervals between 2004/5 and 2006/7 (years 1, 2, and 3 of primary school); wave 4 was conducted 2 years later (year 5 of primary school). Each wave comprised data collection from the primary caregiver, the child, and the teacher. Computer-assisted face-to-face parent interviews were usually conducted at the parent's home. Computer-assisted personal face-to-face child interviews were mostly conducted in the schools. Teacher assessments consisted of one-page paper-and-pencil questionnaires. The median dates for the parent interviews were Oct 2004, Sept 2005, Sept 2006, and Oct 2008. The median dates for the child interviews were March 2005, Dec 2005, Nov 2006 and Jan 2009. The median dates for the teacher assessments were March 2005, Dec 2005, Dec 2006 and August 2008. On average, the baseline assessment was conducted 6 months before the intervention. The post measures were taken 5 months after the parent trainings, and follow-up data were collected 17 (follow-up 1) and 30 (follow-up 2) months after the intervention.

Parents were offered an incentive (about $25) for participation in the study. 1240 parents (74% of the target sample) agreed to participate in the study at wave 1 (baseline). At the post assessment, the retention rates were 95% (parent interviews), 97% (child interviews), and 96% (teacher assessments). At the first follow-up retention, rates were 95% (parent interviews), 96% (child interviews), and 94% (teacher assessments). At the second follow-up assessment, the retention rates were 86% (parent interviews), 83% (child interviews), and 92% (teacher assessments).

At baseline, the mean age of the target child was 7.03 years (SD=0.40) and 48.1% of the children were female. 77.4% lived in households with both biological parents. The sample was highly heterogeneous with 46.1% of the children living in households where both parents had an immigrant background.

Methods

When implemented as a universal prevention strategy, parent training programs often suffer from low participation rates. Thus, studies generally find that only around 15–30% of the target population enroll for program participation, and that often only about 50% of those who have enrolled effectively fully comply with the intervention (Dumas et al. 2007; Dumka et al. 1997; Haggerty et al. 2002; Heinrichs et al. 2005; Morawska and Sanders 2006; Spoth et al. 2000). Such low exposure rates mean that treatment effects become highly diluted amongst the intended target group, and that an intention-to treat analysis of a randomized experiment yields results that are of limited value. In such cases it often becomes desirable to estimate treatment effects on those who effectively received the treatment. In the context of a randomized experiment, this means that we try to establish whether the outcomes of those who accepted the treatment in the treatment condition (the compliers) differ from those in the control condition who were similar in all respects except for the receipt of the treatment. In what follows we propose *propensity score matching* as a strategy for modeling self-selection into treatment, identifying an equivalent subgroup amongst the participants in the control condition, and estimating unbiased treatment effects for the treated (Guo and Fraser 2010).

Propensity Score Matching

Conceptually, propensity score matching is based on the idea of counterfactuals: Generally, counterfactuals are thought-models that ask: What would an outcome be if a

presumed cause (i.e. the treatment) was not present, but everything else could be held constant? Propensity score matching hence aims to find observations in a pool of non-treated subjects that are (on average) undistinguishable from the treated subjects on as many criteria as possible with the exception of the treatment itself. Following the important work by Rosenbaum and Rubin (1983, 1985; also see Rubin and Thomas 1996), propensity score matching has become an increasingly popular approach to estimate causal effects.

The propensity score is the conditional probability of receiving the treatment rather than the control given the observed covariates (Rosenbaum and Rubin 1983). In the current context, the propensity score is the conditional probability of full exposure to the Triple P intervention, given the observed covariates, namely household demographic characteristics, child and family characteristics, and baseline measures of all outcome variables. If two households have the same propensity score given observed covariates, say a $p=.20$ chance of full exposure to Triple P, then these observed covariates will be of no further use in predicting which of these two households will have received full exposure to Triple P. Thus, for these two households, there will be no systematic tendency for the observed covariates to be different for the Triple P exposed and non-exposed. We note, however, that there may still be differences in unmeasured covariates between the exposed and unexposed that may bias the treatment effect estimate.

Propensity score matching is a three-stage process (Guo et al. 2006). The first stage entails *estimating the propensity score*, which is the conditional probability of receiving treatment conditional upon observed covariates. This probability is found by regressing membership in the treated versus untreated group on a set of observed covariates typically by means of a logit or probit regression (D'Agostino 1998). The second stage is the *matching of the treated subjects to the non-treated subjects* in such a way that the two groups are equivalent on all covariates included in the propensity score. In general this entails either matching treated and untreated individuals with similar propensity scores or the re-weighting of the observations in the control group. Various algorithms are available for the matching, including Mahalanobis metric matching, nearest neighbor matching with and without replacement, kernel matching and local linear regression. Guo et al. (2006), Caliendo and Kopeinig (2005), and Becker and Ichino (2002) provide overviews of the advantages and disadvantages of various matching algorithms.

If matching has been successful, the third stage consists of estimating treatment effects based on the balanced treatment and control groups. Strategies may comprise straightforward *t*-tests of mean differences in the outcomes between the treated and the untreated or in multivariate analyses such as generalized linear modeling, survival analysis, or structural equation modeling (Guo et al. 2006; Guo and Fraser 2010).

Simulation studies (Rubin and Thomas 1996) and methodological assessments suggest that propensity score matching can be a powerful tool to estimate unbiased treatment effects (Dehejia and Wahba 2002; Diaz and Handa 2006). However, the adequacy and usefulness of propensity score matching depends on a number of factors. The two most important criteria relate to a sufficient overlap of the propensity to receive treatment in the treated and the control group (Smith and Todd 2005) and a set of high-quality covariates, measured before the intervention, which represent processes associated with selection bias (Morgan and Harding 2006).

Propensity score matching is commonly used for identifying treatment effects in *non-randomized studies*. However, the logic of propensity score matching also applies when allocation to treatment was random, but only a fraction of those in the treatment condition take up the treatment. In this case, the goal of propensity score matching aims at identifying a subgroup of observations in the control condition, whose probability of treatment is equal to those who accepted treatment in the treatment condition.

Defining the Comparison Groups

In the following analyses we compare the *treated in the treatment condition* with matched observations amongst the *untreated in the control condition*. The process of defining the two groups is illustrated in Fig. 1.

It shows that the families of 856 children were allocated to the control condition while the families of 819 children were in the Triple P condition.

In the treatment condition, 69.4% of the parents ($N=568$) participated in wave 1 of the longitudinal study. Amongst these study participants, $N=235$ enrolled in one of the Triple P courses. Parents of 144 children completed all four Triple P sessions and were defined as treatment compliers. Sixty-two parents dropped out of the program prematurely. A comparison of the treatment drop-outs with the compliers reveals significant differences, which justify their exclusion from subsequent analyses. Thus, only 28.4% of drop-outs used the telephone support in comparison to 68.0% of the compliers (χ^2 ($N=220$), $1=29.67$; $p<.001$). Also, about 3–4 months after the program, drop-outs used fewer Triple P techniques than compliers (5.4 vs. 7.4 out of 13 techniques, $F=13.76$; $p<.001$). Furthermore, treatment drop-outs were less likely to report that they were satisfied with the program or that they would recommend it. Thus drop-outs were not only were exposed to fewer program elements but were also less engaged with the program contents.

In the control condition, 672 parents participated in wave 1 of the longitudinal study (78% of the target group).

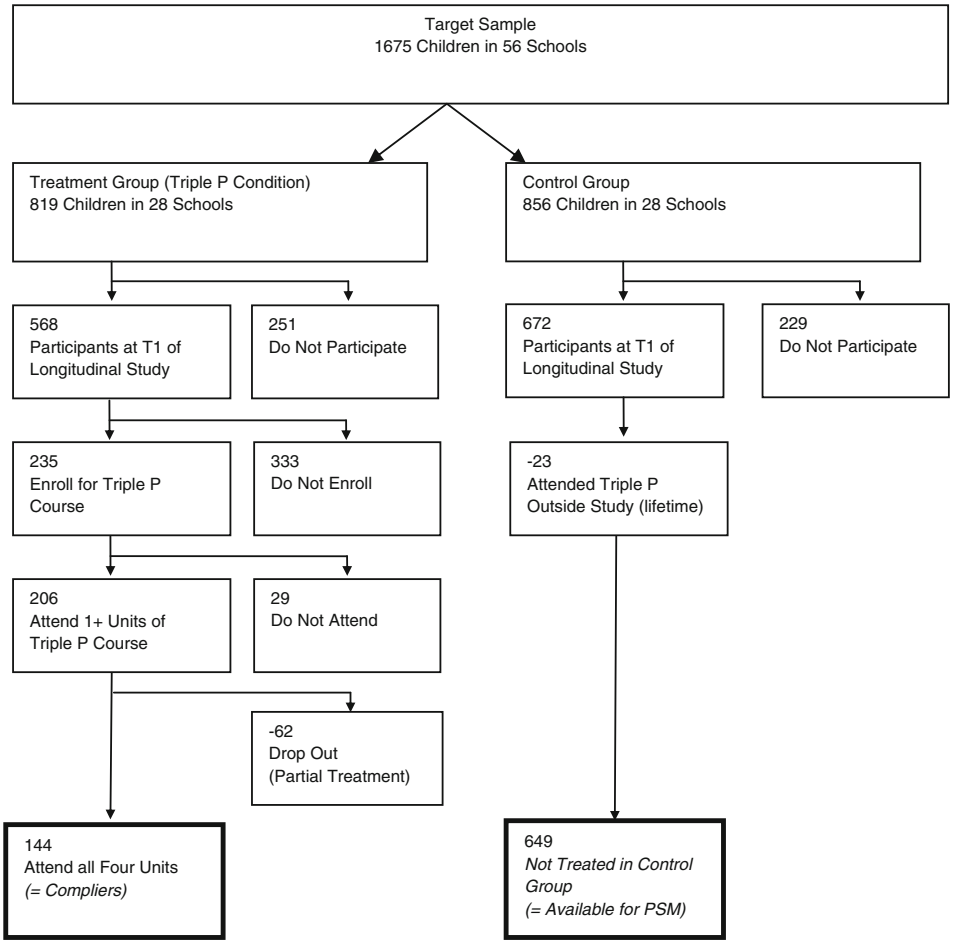

Fig. 1 Definition of the treated in the treatment condition and untreated in the control condition – flow diagram. Note: Groups used for propensity score matching in bold

Note: Groups used for propensity score matching in bold.

Twenty-three parents reported that they had attended a regular (i.e., non-experimental) Triple P program in any of the years preceding the study. They were excluded from further analyses. This leaves 649 untreated parents in the control condition that were available for propensity score matching.

We note that no parent in the control condition enrolled for any of the experimental courses. Also, there was no evidence suggesting that parents in the control condition increased their use of alternative parent-training programs. There was hence no evidence of spillover between treatment conditions, an important component of the stable unit treatment value assumption (SUTVA).

Outcome Measures

Parent training programs are designed to elicit desirable change in parenting behavior, which in turn reduces problematic child behavior (Barlow and Stewart-Brown 2000; Nixon 2002; Serketich and Dumas 1996; Webster-Stratton and Taylor 2001). The current study therefore includes measures of parenting practices as well as of child problem behavior as the targeted outcomes.

The *Alabama Parenting Questionnaire* (APQ) by Shelton et al. (1996) was used to assess parenting practices. The APQ comprises five subscales, namely parental involvement, positive parenting, poor monitoring, erratic discipline, and corporal punishment. It was administered to the primary caregiver in all four waves. The sequence of items was randomized in each interview (using CAPI). Scale reliabilities across the four waves were for parental involvement (10 items) Cronbach's α=.64–.72; positive parenting (5 items) Cronbach's α=.56–.68; poor monitoring (10 items) Cronbach's α=.64–.73; erratic discipline (6 items) Cronbach's α=.52–.58; corporal punishment (3 items) Cronbach's α=.57–.65. Reliabilities are lower than those reported in other studies using the APQ (e.g., Clerkin et al.

2007; Essau et al. 2006; Shelton et al. 1996), possibly because of the heterogeneity of the sample.

Child problem behavior was assessed with the *Social Behavior Questionnaire (SBQ)* developed by Tremblay et al. (1991). It has variously been shown to be change sensitive (e.g., Lacourse et al. 2002; Lösel et al. 2006; Vitaro and Tremblay 1994). The SBQ was used to distinguish five subdimensions, namely *prosocial behavior, internalizing problems, impulsivity and attention deficits*, *non-aggressive conduct problems,* and *aggressive behavior*. In waves 1 and 3, the full version was administered to all respondents. In wave 2, the subdimensions for internalizing behavior and attention deficits were not included in the parent and the child versions. In wave 4, the child version only comprised measures for prosocial and aggressive behavior. In the parent and the child versions the question sequence was randomized. In contrast, the teacher version was a paper-and-pencil assessment with a set question order. In the parent and teacher versions a 5-level Likert scale response format was offered. In the child interviews, drawings illustrating the behavior were presented and children chose between a yes or no option. Across the four waves the reliabilities for the child social behavior subscales were Cronbach's α=.86–.96 in the teacher assessments, .68–.84 in the parent interviews, and .58–.73 in the child interviews.

Missing Values

There were two types of missing data in the present dataset; namely, data missing due to non-response to individual items and missing data due to attrition over the four waves. Across the full dataset the proportion of missing values was 4.1%. The proportion of missing values was 1.1% in wave 1 (baseline), 4.0% in wave 2, 5.2% in wave 3, and 12.4% in wave 4. Little's MCAR test suggested that missing values were not missing completely at random (MCAR=8626.6; df=7820; p<0.001). We therefore used the EM algorithm of SPSS V 18.0 to impute missing data.

Covariates for the Matching Procedure

The goal of propensity score matching is to balance the treatment and the control group on measured covariates that may either be related to the outcome or to the likelihood of treatment exposure (Brookhart et al. 2006). The method therefore depends on the availability of rich data, measured before the intervention and preferably coming from different informants, that represent covariates associated with self-selection into treatment or outcome (Haviland et al. 2007). In this study, 54 covariates were included in the logit models used to estimate propensity scores (see Table 1). All covariates were measured at T1 and were therefore not influenced by the treatment, which was administered about 6 months after the baseline assessment.

Twenty-nine covariates had either been found to predict program participation in this study (Eisner et al. 2011), had been identified as predictors of program participation in other studies; or represented developmental risk factors associated with child problem behavior. Six variables measure child characteristics (sex, age, special needs class, intellectual developmental delay, birth complications, and school performance). Thirteen variables measure aspects of the families' situation including family structure, the socio-demographic background, family functioning, and neighborhood integration. Nine variables distinguish major ethnic-immigrant groups so that propensity score matching balances the treatment and control groups on detailed immigrant background characteristics. Also, one variable measures allocation to the PATHS condition. Inclusion of this variable is conceptually important because it balances the groups in respect of the school-based intervention, which started after wave 2 of the study. Successful balancing on this variable results in maintaining the orthogonal structure of the two interventions, meaning the subsequent analyses of Triple P effects are not influenced by the PATHS intervention. Four variables represent socio-demographic characteristics of the 54 schools included in the study such as average household income and the percentage of families with immigrant background.

Finally, 20 variables represent the baseline measures of all outcome variables. These comprise the five parenting practices measured by the Alabama Parenting Questionnaire and the five child behavior dimensions, each measured from the teacher, the parent, and the child's own perspective.

Results

Propensity Score Matching

Propensity score matching was conducted using the STATA module *psmatch2* developed by Leuven and Sianesi (2003). In a first step, the conditional probability of receiving treatment given the included 54 covariates (i.e., the propensity scores) were computed. The logit models used for estimating the propensity scores were successful in modeling selection into treatment. The model for deriving propensity scores had a likelihood ratio chi-square of 144.63 (df=54; p<0.001; pseudo R^2=19.3%).

Matching was performed with the *nearest neighbor matching algorithm*. In this approach the individual in the control condition that is closest to the propensity score of a treated individual is chosen as a matching partner. Nearest

Table 1 54 covariates included in the propensity score matching

Variable	Value range
PATHS	Dummy, 1 = yes (allocated to PATHS treatment condition)
Child sex	Dummy, 1 = female
Child age_young	Dummy, 1 = below youngest regular school entry age
Child age_old	Dummy, 1 = above highest regular school entry age
Small class	Dummy, 1 = small (special needs) class
Intellectual developmental delay	Dummy, 1 = yes
Birth complications	Dummy, 1 = yes
School performance	Mean score of performance in mathematics and language skills, teacher assessed, 5-point Likert Scale
Mother's age	Birth year of biological mother
Alcohol use during pregnancy	Dummy, 1 = yes
Post-natal depression	Dummy, 1 = yes
Single parent	Dummy, 1 = yes
Child in external child care	Dummy, 1 = yes (more than 3 days per week before age 7)
Dual earner family	Dummy, 1 = yes (both PC's employed 50% or more)
Number of siblings	Dummy, 1 = two or more siblings (living in same household)
Parenting values	Mean score of seven items measuring traditional parenting values
Previous use of parenting services	Dummy, 1 = yes (any of 34 services used before baseline assessment)
PC mother language	0 = German, 1 = Albanian, Turkish, Portuguese (Languages of Triple P courses), 2 = other Non-German.
PC migration background (6 variables)	Former Yugoslavia, Mediterranean countries, Asian, African and Near Eastern, affluent Western, Latin American.
Occupational prestige	Mean ISEI (International Socio-Economic Index of occupational status) score for both primary caregivers
Unemployment	Dummy, 1 = yes (at the time of the baseline assessment)
Neighborhood cohesion	Mean score of 5 items, range=0 to 4.
Neighborhood networks	Mean score of 5 items, range=0 to 4.
Parenting practices (5 variables)	Mean score for each APQ subdimension, range=0 to 4.
Social problem behavior—teacher rated (5 variables)	Mean score for each SBQ subdimension, range=0 to 4.
Social problem behavior—parent rated (5 variables)	Mean score for each SBQ subdimension, range=0 to 4.
Social problem behavior—child rated (5 variables)	Mean score for each SBQ subdimension, range=0 to 1.
School-level income	Mean parental income, by school
School-level % in small class	% children in special needs classes, by school
School-level % immigrant	% children with immigrant background, by school
School-level network density	Mean score of parental networks by school

neighbor matching can be performed with and without replacement. "With replacement" means that individuals in the control group can be used more than once as a match. This results in improved balance, but entails increased variance of the estimator as fewer distinct observations are used to construct the counterfactual (Smith and Todd 2005). In this study nearest neighbor matching was performed *without replacement*. Nearest-neighbor matching without replacement is a 'greedy' algorithm whereby observations in the pool of controls are no longer available one they are matched to a treated observation. The data set was therefore ordered in random sequence before the matching procedure. Nearest neighbor matching is based on finding matches with the smallest distance (in terms of the propensity score) between an observation in the treated group and an observation in the control group.

The counterfactual approach of propensity score matching is predicated on the idea that individuals can be found in the control condition that have propensity scores close to those of the treated individuals. Only for these individuals can a treatment effect be established that assumes all other (measured) variables are balanced. If this common support condition fails, matching cannot be performed (Caliendo and Kopeinig 2005). One form to impose common support is to impose a maximum propensity score difference (a *caliper*) during the matching process (Rosenbaum and

Rubin 1985). In this study, a caliper of 0.125 standard deviations of the propensity score was used, which is a more narrow caliper than the 0.25 suggested by Rosenbaum and Rubin (1985). Within this caliper, adequate matches could be found for all 144 treated individuals.

Furthermore, a decision had to be made on the number of matches sought for each treated individual. We decided to use 1-to-1 matching. The decision was based on the argument that within a randomized experiment with equally sized arms, the expected number of equivalent matches in the control condition corresponds to those who complied with the treatment in the treatment condition.

Table 2 summarizes the results of the matching procedure.

An important tool to assess whether covariate balance has been achieved is the *standardized absolute bias,* which is calculated as

$$\text{Absolute Bias} = 100 * \frac{\bar{x}_{treated} - \bar{x}_{control}}{\sqrt{\frac{s^2_{treated} + s^2_{control}}{2}}}$$

where $x_{treated}$ and $x_{control}$ are the means of a given covariate for the treated and the control condition, respectively. Likewise, $s_{treated}^2$ and $s_{control}^2$ are the respective standard deviations of the given covariate. Rosenbaum and Rubin (1985) have suggested that differences greater than 20% be regarded as unacceptable. Following the recommendations by Haviland et al. (2007), we also show the overall absolute bias of the propensity score as a summary measure of matching success.

Before matching, the mean absolute bias across the 54 variables was 15.71 (*SD*=10.87) and 18 variables had an absolute bias of>20. The absolute bias of the logit propensity score as an overall summary measure of imbalance was 116.08 ($t(791)=14.08$, $p<0.001$).

After matching, the absolute mean bias across the 54 variables was 5.44 (*SD*=4.87). No variable had an absolute bias larger than 20 and no difference between the control and the treatment condition was statistically significant. Also, the absolute bias of the logit propensity score as an overall summary measure of imbalance was 12.30 ($t(288)=1.04$, *n.s.*). Both the reduction in the mean absolute bias across all variables (−65.4%) and the reduction in the bias of the propensity score (−89.5%) suggest that matching was successful. In other words, the 144 matched observations selected from the untreated members of the control group are equivalent to the treated families on measures of child background, family structure, ethnic composition, structural properties of the school context, and all baseline measures of problem behavior.

We also examined 43 variables that had not been used to calculate the propensity scores and assessed their equivalence between the matched groups. Variables considered include 4 teacher-reported measures on the child's social role in the classroom, 10 measures of observer-rated child behavior (e.g., impulsivity, restlessness, attention problems, resistance, aggression), 3 measures taken from the child interviews (sensation seeking, emotion recognition, and sociometric status in the class), 17 parent-reported routine activities of the children at wave 1 as well as 9 parent-reported life events (e.g., moving home, unemployment, death of a family member). Three out of 43 variables were found to differ between the two groups at $p<.10$ and none was found to differ at $p<.05$. This is within the range of associations that can be expected by chance, and suggests that the matching procedure also achieved equivalence for measured variables not included in the propensity matching. However, one cannot exclude the possibility that the groups remain imbalanced on some unmeasured variables.

Treatment Effects

After successful matching several methods can be used to estimate treatment effects. The simplest strategy is to use differences in the post measures as measures of treatment effects. However, several studies suggest that potential bias can be further reduced by using a regression-based approach, where the baseline measure of the outcome is used as a statistical control (Oakes and Feldman 2001; Onur 2006), i.e.

$$Y = \propto + \beta_1 X + \beta_2 T + \varepsilon,$$

where Y is the post-score of an outcome variable, α is the estimated intercept, X is the pretest score of the same variable, and T is a (0,1) indicator for treatment or control

Table 2 Summary statistics of matching success

	Before matching	After matching
Mean absolute bias (*SD*)	15.71 (10.87)	5.44 (4.87)
Maximum absolute bias	39.99	18.61
Absolute bias of logit propensity score	116.08 (t =14.08, $p<0.001$)	12.30 (t=1.04, *n.s.*)
N variables with absolute bias>20	18	0
N variables with difference sig ($p<0.05$)	20	0

1-to-1 Nearest Neighbor Matching, Caliper=0.125, without replacement

group. Treatment effects were hence computed as difference in the outcome between treatment and control groups conditional on pre-test score, using maximum likelihood estimates. Because the data come from a cluster-randomized experiment, we report robust standard errors adjusted for the 56 clusters, using the respective algorithm in STATA. Cohen's *d* effect sizes were computed to assess the standardized size of intervention effects. Standardized effects sizes are coded such that positive values correspond to desirable effects of the intervention. Results are reported in Table 3 (see Tables 4 and 5 for means and standard deviations in the matched and the full samples).

We first examine findings for the five dimensions of parenting behavior. Findings suggest that attendance of the Triple P program did not result in a statistically significant change on any of the five subdimensions of the APQ (parental involvement, positive parenting, parental supervision, erratic parenting, and corporal punishment). No differences were found at any of the three post-intervention assessments. In other words: The self-reported parenting practices amongst program compliers did not differ significantly from those parents in the control condition, who were statistically equivalent on over 80 background characteristics.

In a second step, we examine the effects of Triple P on measures of child problem behavior. Note that internalizing behaviors and ADHD symptoms were not measured in the post-assessment, due to time constraints in that wave of parent interviews. Findings suggest that primary caregivers who completed the Triple P program did not perceive a statistically significant improvement in the child's behavior on any of the five behavior sub-dimensions. This finding holds both for the post measures, the first follow-up, and the second follow-up measures.

Considering the *teacher-assessed* child behaviors we found no significant effects for prosocial behavior, ADHD-related problems, non-aggressive conduct problems and aggressive behavior. In contrast, internalizing problems were perceived by the teachers as developing worse amongst children whose parents attended the program in

Table 3 Effects of Triple P on parenting and child behavior outcomes at post, follow-up 1 and follow-up 2

	Post (5 months)		Follow-up 1 (17 months)		Follow-up 2 (30 months)	
Outcome	B (SE)	Cohen's *d*	B (SE)	Cohen's *d*	B (SE)	Cohen's *d*
Parenting, parent report						
Involvement	0.001 (0.036)	0.00	0.035 (0.033)	0.12	0.071 (0.037)	0.22
Positive parenting	0.005 (0.042)	0.01	−0.065 (0.056)	−0.14	0.029 (0.055)	0.06
Parental supervision	0.004 (0.039)	−0.02	0.021 (0.038)	−0.07	0.016 (0.041)	−0.06
Erratic discipline	0.038 (0.042)	−0.11	−0.027 (0.051)	0.06	−0.075 (0.046)	0.19
Corporal punishment	−0.014 (0.033)	0.05	−0.027 (0.033)	0.10	0.021 (0.033)	−0.08
Child behavior, parent report						
Prosocial	−0.028 (0.038)	0.08	−0.015 (0.043)	0.04	−0.005 (0.055)	−0.01
Internalizing	N/A	N/A	−0.040 (0.056)	0.08	0.018 (0.060)	−0.04
ADHD	N/A	N/A	−0.044 (0.050)	0.10	0.035 (0.072)	−0.06
Non-aggr CD	0.026 (0.033)	−0.09	0.028 (0.030)	−0.11	0.022 (0.035)	0.07
Aggression	−0.012 (0.031)	0.04	−0.001 (0.031)	0.00	−0.005 (0.032)	0.02
Child behavior, teacher report						
Prosocial	−0.005 (0.032)	−0.02	−0.043 (0.094)	−0.05	−0.12 (0.111)	0.17
Internalizing	0.206* (0.100)	−0.24	0.220* (0.089)	−0.29	−0.008 (0.142)	0.01
ADHD	0.147 (0.083)	−0.21	0.127 (0.091)	−0.17	−0.002 (0.117)	−0.00
Non-aggr CD	0.089 (0.054)	−0.19	0.103 (0.062)	−0.20	0.018 (0.055)	−0.04
Aggression	0.047 (0.058)	−0.10	0.037 (0.067)	−0.07	0.056 (0.086)	−0.08
Child behavior, child self-report						
Prosocial	0.028 (0.016)	0.20	−0.018 (0.016	−0.13	−0.005 (0.016)	−0.04
Internalizing	N/A	N/A	−0.046 (0.028)	0.19	N/A	N/A
ADHD	N/A	N/A	−0.020 (0.018)	0.13	N/A	N/A
Non-aggr CD	−0.017 (0.020)	0.10	0.005 (0.018)	−0.04	N/A	N/A
Aggression	−0.023 (0.019)	0.15	−0.008 (0.014)	0.06	−0.026 (0.018)	0.17

* $p<.05$

comparison to non-participants. The effect size is $d=-0.24$ ($p<0.05$) at the post assessment and $d=-0.29$ ($p<0.05$) at the first follow up. However, in the second follow-up assessment, 3 years after the intervention, no difference was found between the children of treated and the untreated parents ($d=-0.01$, *n.s.*). Finally, regarding the children's self-reported behaviors, results show no significant effects for any of the behavioral domains at either of the assessment waves.

Discussion

This study found no effects of a parent training program on five dimensions of parent-reported parenting practices, and on parent-reported and child self-reported problem behaviors. For teachers-assessed problem behaviors, no effects were found for four out of five subdimensions. A small adverse significant effect was observed for teacher assessed internalizing problems at the post and the first follow-up assessments. The effect did not persist at the second follow-up assessment. Given the large number of tested effects, and the small size of the effects it is probably safe to conclude that the parent training program Triple P did not have effects in either direction.

These findings diverge from those reported in previous studies conducted by the program developer or local license holders on Triple P as a universal prevention strategy (McTaggart and Sanders 2003; Heinrichs et al. 2006) and the overall assessment in the meta-analysis by Nowak and Heinrichs (2008). However, they are similar to two other independent trials: In particular, de Graaf et al. (2009) found no effects of Primary Care Triple P on any dimension of child problem behavior in a trial in the Netherlands, although a positive effect was found on lax parenting. Similarly, McConnell et al. (2011) found no effects of Primary Care Triple P on parent, child, and family outcomes in an independent trial conducted in Canada. They thus add to the literature that finds no or greatly reduced effects in replication studies with no involvement of the program developer (Petrosino and Soydan 2005). When independent replications fail to corroborate findings reported in developer-led studies, one should first examine whether the lack of effects can be attributed to shortcomings in the replication study.

To assess this possibility, the present findings are best compared with the two studies that examined group-based level 4 Triple P as a universal intervention, that were similar in study design, and that were conducted by the program developer or local license holders. In the study by McTaggart and Sanders (2003), 25 schools in Brisbane (out of 78 contacted schools) were randomly allocated to a control and an intervention condition, and group-based level 4 Triple P was offered to parents of year one classes. The study collected data on teacher-assessed problem behavior and found a reduction both in the Sutter-Eyberg Student Behavior Inventory (SESBI; Eyberg and Ross 1978) problem and intensity scales. The effect size for the intensity score at the post-measure was about $d=0.14$. Between-group follow-up effect sizes are not available for this study. In the study by Heinrichs et al. (2006), 17 pre-school day-care centers (out of 33 contacted centers) in Braunschweig (Germany) were allocated to treatment and control conditions, and parents were offered group-based level 4 Triple P. The study relied on parent assessments. It found significant reductions in parenting problems and child problem behavior according to the mothers' reports, but no effects according to the fathers' assessments. For the mother-assessed child problem behavior score, the authors report an effect size of $d=0.38$ at post and $d=0.32$ at follow up. These results were found after the group of non-compliant participants in the treatment condition had been retrospectively re-allocated to the control condition (Heinrichs et al. 2006, p. 87). No results are reported for the treatment effects according to initial treatment assignment.

In comparing our findings with those in the other two studies, we first examined evidence for differences in implementation quality. As mentioned above, 27% of the parents in the treatment condition in the Zurich study attended at least one session. The respective participation rates were 11% in Brisbane (McTaggart and Sanders 2003, p. 5) and 24% in Braunschweig (Heinrichs et al. 2006), suggesting that the Zurich study achieved a rather high participation rate. Furthermore, the Braunschweig study reports customer satisfaction scores. In Braunschweig, 91% of the mothers were satisfied with the program and 94% found the program useful (Heinrichs et al. 2006, p. 88). In Zurich, the respective rates were almost identical with 89% being satisfied and 91% finding the program useful. Finally, in all three studies experienced and licensed facilitators provided the courses, using standardized treatment manuals, and similar supervision arrangements were put in place for the facilitators. Altogether, thus, these data do not suggest that discrepancies in implementation quality were responsible for the observed differences in treatment effects.

A second possibility is that the target group in the Zurich study was not receptive to the intervention. However, the introduction of Triple P in Zurich was based on a comprehensive needs assessment, which indicated that a universal parent training was not yet available and would fit well into the overall public health strategy of the city. Also, the comparatively high recruitment rate suggests that parents were sympathetic to the program. Finally, the urban contexts of Brisbane, Braunschweig and Zurich are comparable in respect of city size, per capita income, family structure, and life-style and value

orientations. However, we acknowledge that the sample in the current study was recruited from the general population, in which only a small fraction can be expected to be at risk for dysfunctional parenting practices, and that the existing provision of parent support in Zurich is comparatively good. This may partly account for the lack of positive results in the present study.

Furthermore, the discrepancies may be due to the different measurement instruments. Thus, McTaggart and Sanders (2003) relied on the SESBI (Eyberg and Ross 1978) to measure child problem behavior; Heinrichs et al. (2006) administered the *Achenbach Child Behavior Checklist* (Achenbach and Edelbrock 1981), while the Zurich study used Tremblay's *Social Behavior Questionnaire* (Tremblay et al. 1991). However, the response scales used in the three instruments are similar and many items are equivalent. Also, all three instruments have been shown to be change-sensitive in intervention studies. Furthermore, we explored whether the lack of effects in this study may be attributed to ceiling or floor effects in the dependent variables. However, while problem behavior measures were – as is to be expected in a population sample – skewed to the right, all indicators included relatively frequent items and had considerable variance, meaning that floor effects are an unlikely reason for the lack of effects.

Finally, the discrepancies could result from differences in the methodological rigor of design and statistical analyses. In this respect we believe that the current study compares favorably with the two aforementioned studies. For example, it is the only study that used a multi-informant approach based on the primary caregiver, the teacher, and the child, while the Brisbane and the Braunschweig studies rely exclusively on teacher or parent reports, respectively. Also, in the Braunschweig and the Brisbane studies, only a fraction of the contacted schools/pre-school institutions participated in the study (17 out of 33 and 25 out of 78, respectively), while in Zurich all contacted schools could be recruited for participation. This limits self-selection and expectancy effects at the level of the participating aggregate units and increases generalizability. Furthermore, the study participation rate in the Braunschweig study was 31% (Heinrichs et al. 2005) in comparison to 74% in the Zurich study, meaning that the latter results are more generalizable to the study population. Also, the Braunschweig study reports high baseline differences in problem behavior scores between the treated and the control condition (e.g., baseline $CBCL_{treated}$ M =33.1, SD=20.1 vs. $CBCL_{control}$ M =26.3, SD=14.0). They suggest problems with the randomization and make it difficult to distinguish treatment effects from mere regression to the mean. In contrast, the treated and the controls in this study were not only balanced on all baseline measures of the core outcome measures, but also on a large number of other background variables.

While we believe that the present study has significant strengths in comparison with the most similar extant studies, we also note important limitations: Thus, as a result of the longitudinal character of the study, post-intervention effects were measured 4–6 months after the delivery of the program. Possible effects immediately following the intervention could therefore not be established. Also, ADHD and internalizing behavior were not measured in wave 2 of the parent interviews and we were not able to determine short-term effects on those dimensions. Further, the propensity score matching approach taken in the present study made it impossible to account for the clustered nature of the data, as methods for conducting propensity score matching with hierarchical data are still in their infancy (Arpino and Mealli 2011).

The findings reported in this study have broader implications. First, they suggest that some modes of delivery of Triple P may be ineffective in some contexts, while we acknowledge that it belongs to the most thoroughly evaluated parent training programs worldwide with an impressive body of findings in support of its efficacy (Nowak and Heinrichs 2008). In particular, our findings suggest that group-based behavioral parent training may not be an effective universal strategy for broad populations of parents of primary school-children. Second, they add to the evidence that findings from experiments with a large influence of the program developers cannot always be generalized. We believe that the comparatively large number of failed replications in field trials is cause for concern. We therefore concur with others (e.g., St. Pierre et al. 2005) that high-quality independent field trials are an essential step towards a better evidence base for effective prevention of child and adolescent problem behaviors. In such independent replications, every possible effort should be made to rule out low implementation quality or study design bias as possible explanations of results. Furthermore, more efforts should be made to better understand the mechanisms that cause the systematic differences in effectiveness found in developer-led studies and in independent replications. Petrosino and Soydan (2005) suggested that discrepancies might either result from implementation failures in independent trials or from a systematic bias, possibly due to conflict of interest, in developer-led studies. It is currently impossible to say which of these possible explanations better accounts for the empirical patterns. We believe that more research on this issue is essential to promote our understanding of how prevention programs can be effective under conditions of routine applications in field settings.

Acknowledgements We wish to acknowledge financial support for the study by the Swiss National Science Foundation, the Jacobs Foundation, the Swiss Federal Office of Public Health, the Canton of Zurich Ministry of Education, and the Julius Baer Foundation. We also thank the anonymous reviewers for helpful comments on earlier versions of this article.

Appendix

Table 4 Means and standard deviations of all dependent variables in the treated and matched untreated subgroups at pre (T1), post (T2), follow-up 1 (T3) and follow-up 2 (T4) assessments

	Compliant in treatment condition (*N*=144)				Matched controls in control condition (*N*=144)			
	T1	T2	T3	T4	T1	T2	T3	T4
A) Parenting, parent report								
Involvement	3.23 (0.37)	3.07 (0.38)	3.07 (0.37)	3.02 (0.37)	3.21 (0.33)	3.06 (0.37)	3.02 (0.34)	2.94 (0.35)
Positive parenting	3.07 (0.45)	3.07 (0.46)	3.02 (0.48)	2.99 (0.51)	3.07 (0.48)	3.07 (0.46)	3.09 (0.49)	2.96 (0.34)
Poor supervision	0.32 (0.31)	0.38 (0.34)	0.40 (0.36)	0.49 (0.42)	0.34 (0.36)	0.39 (0.36)	0.39 (0.36)	0.53 (0.42)
Erratic parenting	1.26 (0.50)	1.19 (0.50)	1.18 (0.52)	1.14 (0.56)	1.28 (0.55)	1.17 (0.51)	1.22 (0.50)	1.23 (0.52)
Corporal punishment	0.36 (0.40)	0.25 (0.36)	0.23 (0.35)	0.19 (0.34)	0.32 (0.43)	0.25 (0.35)	0.24 (0.39)	0.15 (0.32)
B) Child behavior, parent report								
Prosocial	2.53 (0.50	2.58 (0.46)	2.59 (0.50	2.63 (0.52)	2.50 (0.49)	2.59 (0.47)	2.59 (0.49)	2.63 (0.52)
Internalizing	0.75 (0.49)	__[a]	0.88 (0.51)	0.93 (0.57)	0.76 (0.43)	__[a]	0.93 (0.48)	0.91 (0.48)
ADHD	1.24 (0.62)	__[a]	1.31 (0.66)	1.28 (0.69)	1.28 (0.61)	__[a]	1.40 (0.64)	1.28 (0.66)
Non-aggr CD	0.69 (0.36)	0.74 (0.40)	0.70 (0.36)	0.66 (0.42)	0.70 (0.39)	0.73 (0.35)	0.68 (0.39)	0.65 (0.40)
Aggression	0.75 (0.40)	0.77 (0.43)	0.72 (0.41)	0.55 (0.38)	0.78 (0.51)	0.81 (0.47)	0.74 (0.46)	0.57 (0.40)
C) Child behavior, teacher report								
Prosocial	2.20 (0.84)	2.25 (0.80)	2.34 (0.72)	2.16 (0.77)	2.18 (0.90)	2.16 (0.87)	2.36 (0.87)	3.32 (0.79)
Internalizing	0.90 (0.82)	0.90 (0.78)	0.94 (0.77)	0.92 (0.75)	0.93 (0.79)	0.70 (0.65)	0.74 (0.70)	0.93 (0.78)
ADHD	1.14 (0.96)	1.10 (0.95)	1.00 (0.84)	0.99 (0.88)	1.29 (1.07)	1.06 (0.98)	0.96 (0.90)	1.07 (1.01)
Non-aggr CD	0.30 (0.50)	0.33 (0.51)	0.34 (0.48)	0.25 (0.37)	0.39 (0.49)	0.30 (0.45)	0.28 (0.47)	0.27 (0.50)
Aggression	0.59 (0.70)	0.54 (0.63)	0.58 (0.53)	0.55 (0.58)	0.68 (0.72)	0.55 (0.63)	0.58 (0.76)	0.53 (0.69)
D) Child behavior, child self-report								
Prosocial	0.84 (0.16)	0.90 (0.13)	0.90 (0.13)	0.90 (0.13)	0.83 (0.18)	0.87 (0.19)	0.92 (0.16)	0.90 (0.13)
Internalizing	0.41 (0.24)	__[a]	0.36 (0.26)	__[a]	0.41 (0.23)	__[a]	0.41 (0.23)	__[a]
ADHD	0.15 (0.17)	__[a]	0.16 (0.18)	__[a]	0.15 (0.17)	__[a]	0.18 (0.20)	__[a]
Non-aggr CD	0.18 (0.15)	0.19 (0.17)	0.18 (0.17)	__[a]	0.19 (0.17)	0.22 (0.19)	0.18 (0.17)	__[a]
Aggression	0.18 (0.17)	0.15 (0.18)	0.13 (0.15)	0.20 (0.18)	0.19 (0.17)	0.18 (0.18)	0.13 (0.16)	0.23 (0.18)

[a] Not measured in the respective wave.

Table 5 Means and standard deviations in the full sample at pre (T1), post (T2), follow-up 1 (T3) and follow-up 2 (T4) assessments

	Triple P Treatment Condition (*N*=568)				Non-Triple P Control Condition (*N*=672)			
	T1	T2	T3	T4	T1	T2	T3	T4
A) Parenting, parent report								
Involvement	3.21 (0.44)	3.11 (0.42)	3.10 (0.40)	3.04 (0.43)	3.18 (0.41)	3.09 (0.42)	3.06 (0.41)	3.00 (0.41)
Positive Parenting	3.23 (0.52)	3.19 (0.53)	3.19 (0.54)	3.13 (0.57)	3.20 (0.51)	3.14 (0.50)	3.13 (0.52)	3.08 (0.54)
Supervision	0.28 (0.31)	0.32 (0.33)	0.34 (0.33)	0.45 (0.40)	0.34 (0.34)	0.37 (0.35)	0.38 (0.36)	0.48 (0.42)
Erratic Parenting	1.25 (0.55)	1.20 (0.54)	1.24 (0.57)	1.24 (0.57)	1.24 (0.54)	1.22 (0.52)	1.21 (0.51)	1.24 (0.51)
Corporal Punishment	0.48 (0.49)	0.39 (0.45)	0.37 (0.47)	0.28 (0.42)	0.43 (0.49)	0.38 (0.46)	0.34 (0.44)	0.23 (0.41)
B) Child behavior, parent report								
Prosocial	2.58 (0.53)	2.71 (0.51)	2.69 (0.52)	2.74 (0.53)	2.58 (0.52)	2.67 (0.53)	2.66 (0.54)	2.70 (0.54)
Internalizing	0.70 (0.48)	__[a]	0.83 (0.48)	0.89 (0.53)	0.71 (0.45)	__[a]	0.87 (0.49)	0.88 (0.50)
ADHD	1.20 (0.64)	__[a]	1.29 (0.64)	1.28 (0.66)	1.22 (0.65)	__[a]	1.30 (0.70)	1.25 (0.68)
Non-aggr CD	0.58 (0.38)	0.61 (0.41)	0.57 (0.39)	0.57 (0.39)	0.61 (0.38)	0.64 (0.40)	0.59 (0.41)	0.57 (0.39)
Aggression	0.58 (0.39)	0.64 (0.42)	0.63 (0.41)	0.49 (0.33)	0.63 (0.45)	0.68 (0.45)	0.66 (0.44)	0.50 (0.38)

Table 5 (continued)

	Triple P Treatment Condition (N=568)				Non-Triple P Control Condition (N=672)			
	T1	T2	T3	T4	T1	T2	T3	T4
C) Child behavior, teacher report								
Prosocial	2.11 (0.81)	2.26 (0.79)	2.30 (0.79)	2.14 (0.76)	2.26 (0.81)	2.30 (0.83)	2.48 (0.85)	2.24 (0.78)
Internalizing	0.92 (0.78)	0.87 (0.75)	0.91 (0.71)	0.87 (0.71)	0.82 (0.73)	0.71 (0.70)	0.79 (0.74)	0.88 (0.70)
ADHD	1.32 (0.98)	1.21 (0.96)	1.16 (0.89)	1.17 (0.99)	1.18 (1.00)	1.01 (0.97)	1.01 (0.98)	1.07 (0.95)
Non-aggr CD	0.36 (0.52)	0.35 (0.54)	0.39 (0.55)	0.30 (0.47)	0.29 (0.45)	0.26 (0.45)	0.28 (0.47)	0.25 (0.45)
Aggression	0.63 (0.71)	0.60 (0.67)	0.61 (0.60)	0.59 (0.69)	0.54 (0.45)	0.48 (0.57)	0.53 (0.65)	0.50 (0.67)
D) Child behavior, child self-report								
Prosocial	0.83 (0.16)	0.89 (0.13)	0.91 (0.12)	0.90 (0.13)	0.81 (0.19)	0.88 (0.16)	0.91 (0.14)	0.89 (0.14)
Internalizing	0.40 (0.24)	—	0.38 (0.24)	—	0.42 (0.23)	—	0.39 (0.24)	—
ADHD	0.17 (0.19)	—	0.14 (0.18)	—	0.18 (0.18)	—	0.18 (0.19)	—
Non-aggr CD	0.21 (0.18)	0.19 (0.17)	0.16 (0.16)	—	0.22 (0.19)	0.21 (0.17)	0.18 (0.17)	—
Aggression	0.16 (0.16)	0.13 (0.15)	0.12 (0.15)	0.19 (0.18)	0.19 (0.18)	0.15 (0.17)	0.14 (0.16)	0.21 (0.19)

[a] Not measured in the respective wave.

References

Achenbach, T. M., & Edelbrock, C. S. (1981). Behavioral problems and competencies reported by parents of normal and disturbed children aged four through sixteen. *Monographs of the Society for Research in Child Development, 46*, 1–82.

Arpino, B., & Mealli, F. (2011). The specification of the propensity score in multilevel observational studies. *Computational Statistics and Data Analysis, 55*, 1770–1780.

Barlow, J., & Stewart-Brown, S. (2000). Behavior problems and group-based parent education programs. *Journal of Developmental and Behavioral Pediatrics, 21*, 356–370.

Bauer, N. S., Lozano, P., & Rivara, F. P. (2007). The effectiveness of the Olweus Bullying Prevention Program in public middle schools: A controlled trial. *The Journal of Adolescent Health, 40*, 266–274.

Becker, S. O., & Ichino, A. (2002). Estimation of average treatment effects based on propensity scores. *The Stata Journal, 2*, 358–377.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Sturmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology, 163*, 1149–1156.

Caliendo, M., & Kopeinig, S. (2005). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys, 22*, 31–72.

Capaldi, D. M., Chamberlain, P., & Patterson, G. R. (1997). Ineffective discipline and conduct problems in males: Association, late adolescent outcomes, and prevention. *Aggression and Violent Behavior, 2*, 343–353.

Clerkin, S. M., Marks, D. J., Policaro, K. L., & Halperin, J. M. (2007). Psychometric properties of the Alabama Parenting Questionnaire-preschool revision. *Journal of Clinical Child & Adolescent Psychology, 36*, 19–28.

D'Agostino, R. B. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine, 17*, 2265–2281.

de Graaf, I., Onrust, S., Haverman, M., & Janssens, J. (2009). Helping families improve: An evaluation of two primary care approaches to parenting support in the Netherlands. *Infant and Child Development, 18*, 481–501.

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics, 84*, 151–161.

Diaz, J. J., & Handa, S. (2006). An assessment of propensity score matching as a nonexperimental impact estimator: Evidence from Mexico's PROGRESA program. *Journal of Human Resources, XLI*, 319–345.

Dumas, J. E., Nissley-Tsiopinis, J., & Moreland, A. D. (2007). From intent to enrollment, attendance, and participation in preventive parenting groups. *Journal of Child and Family Studies, 16*, 1–26.

Dumka, L. E., Grarza, C. A., Roosa, M. W., & Stoerzinger, H. D. (1997). Recruitment and retention of high-risk families into a preventive parent training intervention. *Journal of Primary Prevention, 18*, 25–37.

Eisner, M., & Ribeaud, D. (2005). A randomised field experiment to prevent violence: The Zurich intervention and prevention project at schools, ZIPPS. *European Journal of Crime, Criminal Law and Criminal Justice, 13*, 27–43.

Eisner, M., & Ribeaud, D. (2007). Conducting a criminological survey in a culturally diverse context: Lessons from the Zurich project on the social development of children. *European Journal of Criminology, 4*, 271–288.

Eisner, M., Meidert, U., Ribeaud, D., & Malti, T. (2011). From enrollment to utilization – stages of parental engagement in a universal parent training program. *Journal of Primary Prevention, 32*, 83–93.

Essau, C. A., Sasagawa, S., & Frick, P. J. (2006). Psychometric properties of the Alabama Parenting Questionnaire. *Journal of Child and Family Studies, 15*, 597–616.

Eyberg, S. M., & Ross, A. W. (1978). Assessment of child behavior problems: The validation of a new inventory. *Journal of Consulting and Clinical Psychology, 7*, 113–116.

Farrington, D. P., & Welsh, B. C. (2007). *Saving children from a life of crime; early risk factors and effective interventions*. Oxford, UK: Oxford University Press.

Friedman, L. S., & Richter, E. D. (2004). Relationship between conflicts of interest and research results. *Journal of General Internal Medicine, 19*, 51–56.

Gottfredson, D., Kumpfer, K., Polizzi-Fox, D., Wilson, D., Puryear, V., Beatty, P., & Vilmenay, M. (2006). The Strengthening Washington D.C. Families project: A randomized effectiveness trial of family-based prevention. *Prevention Science, 7*, 57–74.

Greenberg, M. T., Kusché, C. A., & Mihalic, S. F. (1998). *Blueprints for violence prevention, book ten: Promoting Alternative Thinking Strategies (PATHS)*. Boulder, CO: Center for the Study and Prevention of Violence.

Gross, D., Garvey, C., Julion, W., Fogg, L., Tucker, S., & Mokros, H. (2009). Efficacy of the Chicago Parent Program with low-income African American and Latino parents of young children. *Prevention Science, 10*, 54–65.

Guo, S., & Fraser, M. W. (2010). *Propensity score analysis: Statistical methods and applications*. Thousand Oaks, CA: Sage.

Guo, S., Barth, R. P., & Gibbons, C. (2006). Propensity score matching strategies for evaluating substance abuse services for child welfare clients. *Children and Youth Services Review, 28*, 357–383.

Haggerty, K. P., Fleming, C. B., Lonczak, H. S., Oxford, M. L., Harachi, T. W., & Catalano, R. F. (2002). Predictors of participation in parenting workshops. *Journal of Primary Prevention, 22*, 375–387.

Hallfors, D., Cho, H., Sanchez, V., Khatapoush, S., Kim, H. M., & Bauer, D. (2006). Efficacy vs effectiveness trial results of an indicated "model" substance abuse program: Implications for public health. *American Journal of Public Health, 96*, 2254–2259.

Haviland, A., Nagin, D. S., & Rosenbaum, P. R. (2007). Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychological Methods, 12*, 247–267.

Heinrichs, N., Bertram, H., Kuschel, A., & Hahlweg, K. (2005). Parent recruitment and retention in a universal prevention program for child behavior and emotional problems: Barriers to research and program participation. *Prevention Science, 6*, 275–286.

Heinrichs, N., Hahlweg, K., Bertram, H., Kuschel, A., Naumann, S., & Harstick, S. (2006). Die langfristige Wirksamkeit eines Elterntrainings zur universellen Prävention kindlicher Verhaltensstörungen: Ergebnisse aus Sicht der Mütter und Väter. *Zeitschrift für klinische Psychologie und Psychotherapie, 35*, 82–96.

Hiscock, H., Bayer, J. K., Price, A., Ukoumunne, O. C., Rogers, S., & Wake, M. (2008). Universal parenting programme to prevent early childhood behavioural problems: Cluster randomised trial. *BMJ, 336*, 318–321.

Jenson, J. M., & Dieterich, W. A. (2007). Effects of a skills-based prevention program on bullying and bully victimization among elementary school children. *Prevention Science, 8*, 285–296.

Lacourse, E., Côté, S., Nagin, D. S., Vitaro, F., Brendgen, M., & Tremblay, R. E. (2002). A longitudinal-experimental approach to testing theories of antisocial behavior development. *Development and Psychopathology, 14*, 909–924.

Leuven, E., & Sianesi, B. (2003). PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing. Retrieved from http://ideas.repec.org/c/boc/bocode/s432001.html.

Loeber, R., & Stouthamer-Loeber, M. (1986). Family factors as correlates and predictors of juvenile conduct problems and delinquency. In M. Tonry & N. Morris (Eds.), *Crime and justice* ((pp, Vol. 7, pp. 29–149). Chicago, IL: Chicago University Press.

Lösel, F., Beelmann, A., Stemmler, M., & Jaursch, S. (2006). Probleme des Sozialverhaltens im Vorschulalter: Evaluation des Eltern- und Kindertrainings EFFEKT. *Zeitschrift für klinische Psychologie und Psychotherapie, 35*, 127–139.

Lundahl, B., Risser, H. J., & Lovejoy, M. C. (2006). A meta-analysis of parent training: Moderators and follow-up effects. *Clinical Psychology Review, 26*, 86–104.

Malti, T., Ribeaud, D., & Eisner, M. (2011). The effects of two universal preventive interventions to reduce children's externalizing behavior: A cluster randomized controlled trial. *Journal of Clinical Child and Adolescent Psychology, 40*, 677–692. doi:10.1080/15374416.2011.597084.

Maughan, D. R., Christiansen, E., Jenson, W. R., Olympia, D., & Clark, E. (2005). Behavioral parent training as a treatment for externalizing behaviors and disruptive behavior disorders: A meta-analysis. *School Psychology Review, 34*, 267–286.

McConnell, D., Breitkreuz, R., & Savage, A. (2011). Independent evaluation of the Triple P Positive Parenting Program in family support service settings. *Child & Family Social Work.* doi:10.1111/j.1365-2206.2011.00771.

McTaggart, P., & Sanders, M. R. (2003). The transition to school project: Results from the classroom. *Australian e-Journal for the Advancement of Mental Health, 2*, 1–12.

Morawska, A., & Sanders, M. R. (2006). A review of parental engagement in parenting interventions and strategies to promote it. *Journal of Children's Services, 1*, 29–40.

Morgan, S. L., & Harding, D. J. (2006). Matching estimators of causal effects: Prospects and pitfalls in theory and practice. *Sociological Methods Research, 35*, 3–60.

Nixon, R. D. V. (2002). Treatment of behavior problems in preschoolers: A review of parent training programs. *Clinical Psychology Review, 22*, 525–546.

Nowak, C., & Heinrichs, N. (2008). A comprehensive meta-analysis of Triple P-Positive Parenting Program using hierarchical linear modeling: Effectiveness and moderating variables. *Clinical Child and Family Psychology Review, 11*, 114–144.

Oakes, J. M., & Feldman, H. A. (2001). Statistical power for nonequivalent pretest-posttest designs: The impact of change-score versus ANCOVA models. *Evaluation Review, 25*, 3–28.

Onur, B. (2006). Too much ado about propensity score models? Comparing methods of propensity score matching. *Value in Health, 9*, 377–385.

Perlis, R. H., Perlis, C. S., Wu, Y., Hwang, C., Joseph, M., & Nierenberg, A. A. (2005). Industry sponsorship and financial conflict of interest in the reporting of clinical trials in psychiatry. *American Journal of Psychiatry, 162*, 1957–1960.

Petrosino, A., & Soydan, H. (2005). The impact of program developers as evaluators on criminal recidivism: Results from meta-analyses of experimental and quasi-experimental research. *Journal of Experimental Criminology, 1*, 435–450.

Piquero, A., Farrington, D. P., Welsh, B. C., Tremblay, R. E., & Jennings, W. (2009). Effects of early family/parent training programs on antisocial behavior and delinquency. *Journal of Experimental Criminology, 5*, 83–120.

Reyno, S. M., & McGrath, P. J. (2006). Predictors of parent training efficacy for child externalizing behavior problems - a meta-analytic review. *Journal of Child Psychology and Psychiatry, 47*, 99–111.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41–55.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39*, 33–38.

Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics, 52*, 249–264.

Sanders, M. R. (1999). Triple P-Positive Parenting Program: Towards an empirically validated multilevel parenting and family support strategy for the prevention of behaviour and emotional problems in children. *Clinical Child and Family Psychology Review, 2*, 71–89.

Sanders, M. R., Turner, K. T., & Markie-Dadds, C. (2002). The development and dissemination of the Triple P—Positive Parenting Program: A multilevel, evidence-based system of parenting and family support. *Prevention Science, 3*, 173–189.

Sanders, M. R., Markie-Dadds, C., & Turner, K. T. (2003). Theoretical, scientific and clinical foundations of the Triple P Positive Parenting Program: A population approach to the promotion of

parenting competence. *Parenting Research and Practice Monograph, 1*, 1–21.

Serketich, W. J., & Dumas, J. E. (1996). The effectiveness of behavioral parent training to modify antisocial behavior in children: A meta-analysis. *Behavior Therapy, 27*, 171–186.

Shelton, K. K., Frick, P. J., & Wootton, J. (1996). Assessment of parenting practices in families of elementary school-age children. *Journal of Clinical Child Psychology, 25*, 317–329.

Smith, J. A., & Todd, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics, 125*, 305–353.

Spoth, R. L. (2001). Randomized trial of brief family interventions for general populations: Adolescent substance use outcomes 4 years following baseline. *Journal of Consulting and Clinical Psychology, 69*, 627.

Spoth, R. L., Redmond, C., & Shin, C. (2000). Modeling factors influencing enrollment in family-focused preventive intervention research. *Prevention Science, 1*, 213–225.

Spoth, R. L., Kavanagh, K. A., & Dishion, T. J. (2002). Family-centered preventive intervention science: Toward benefits to larger populations of children, youth, and families. *Prevention Science, 3*, 145–152.

St. Pierre, T. L., Osgood, D. W., Mincemoyer, C. C., Kaltreider, D. L., & Kauh, T. J. (2005). Results of an independent evaluation of project ALERT delivered in schools by cooperative extension. *Prevention Science, 6*, 305–317.

Tremblay, R. E., Loeber, R., Gagnon, C., Charlebois, P., Larivée, S., & LeBlanc, M. (1991). Disruptive boys with stable and unstable high fighting behavior patterns during junior elementary school. *Journal of Abnormal Child Psychology, 19*, 285–300.

Vitaro, F., & Tremblay, R. E. (1994). Impact of a prevention program on aggressive children's friendships and social adjustment. *Journal of Abnormal Child Psychology, 22*, 457–475.

Webster-Stratton, C., & Taylor, T. (2001). Nipping early risk factors in the bud: Preventing substance abuse, delinquency, and violence in adolescence through interventions targeted at young children (0–8 years). *Prevention Science, 2*, 165–192.